Predicting landfalling hurricane numbers from sea surface temperature: theoretical comparisons of direct and indirect approaches

Stephen Jewson (RMS)*
Thomas Laepple (AWI)
Kechi Nzerem (RMS)
Jeremy Penzer (LSE)

February 2, 2008

Abstract

We consider two ways that one might convert a prediction of sea surface temperature (SST) into a prediction of landfalling hurricane numbers. First, one might regress historical numbers of landfalling hurricanes onto historical SSTs, and use the fitted regression relation to predict future landfalling hurricane numbers given predicted SSTs. We call this the direct approach. Second, one might regress *basin* hurricane numbers onto historical SSTs, estimate the proportion of basin hurricanes that make landfall, and use the fitted regression relation and estimated proportion to predict future landfalling hurricane numbers. We call this the *indirect* approach. Which of these two methods is likely to work better? We answer this question for two simple models. The first model is reasonably realistic, but we have to resort to using simulations to answer the question in the context of this model. The second model is less realistic, but allows us to derive a general analytical result.

1 Introduction

There is a great need to predict the distribution of the number of hurricanes that might make landfall in the US in the next few years. Such predictions are of use to all the entities that are affected by hurricanes, ranging from local and national governments to insurance and reinsurance companies. How, then, should we make such predictions? There is no obvious best method. For instance, one might consider making a prediction based on time-series analysis of the time-series of historical landfalling hurricane numbers; one might consider making a prediction of basin hurricane numbers using time-series analysis, and convert that prediction to a prediction of landfalling hurricane numbers; one might consider trying to predict SSTs first, and convert that prediction to a prediction of landfalling numbers; or one might try and use output from a numerical model of the climate system. All of these are valid approaches, and each has their own pros and cons.

In this article, we consider the idea of first predicting SST and then predicting hurricane numbers given a prediction of SST. There are two obvious flavours of this. The first is what we will call the 'direct' (or 'one-step') method, in which one regresses historical numbers of landfalling hurricanes directly onto historical SSTs, and uses the fitted regression relation to convert a prediction of future SSTs into a prediction of future hurricane numbers. The second is what we will call the 'indirect' (or 'two-step') method, in which one regresses *basin* hurricane numbers onto historical SSTs, predicts basin numbers, and then predicts landfalling numbers from basin numbers. In the simplest version of the indirect method one might predict landfalling numbers as a constant proportion of the number of basin hurricanes, where this proportion is estimated using historical data.

* *Correspondence email*: stephen.jewson@rms.com

Consideration of the direct and indirect SST-based methods motivates the question: at a theoretical level, which of these two methods is likely to work best? This is a statistical question about the properties of regression and proportion models. We consider this abstract question in the context of two simple models. The first model is the more realistic of the two. It uses observed SSTs, models the mean number of hurricanes in the basin as a linear function of SST, and models each basin hurricane as having a constant probability of making landfall. We run simulations that allow us to directly compare the performance of the direct and indirect methods in the context of this model. The second model is less realistic, but allows us to derive a general analytical result for the relative performance of the direct and indirect methods. In this model we represent SST, basin and landfalling hurricane numbers as being normally distributed and linearly related.

We don't think the answer as to which of the direct or indirect methods is better is *a priori* obvious. On the one hand, the direct method has fewer parameters to estimate, which might work in its favour. On the other hand, the indirect method allows us to use more data by incorporating the basin hurricane numbers into the analysis.

Section 2 describes the methods used in the simulation study, and section 3 describes the results from that study. In section 4 we derive general analytic results for the linear-normal model. Finally in section 5 we discuss our results.

2 Simulation-based analysis: methods

For our simulation study, we compare the direct and indirect methods described above as follows.

2.1 Generating artificial basin hurricane numbers

First, we simulate 10,000 sets of artificial basin hurricane numbers for the period 1950-2005, giving a total of 10,000 x 56 = 560,000 years of simulated hurricane numbers. These numbers are created by sampling from poisson distributions with mean given by:

$$\lambda = \alpha + \beta S \tag{1}$$

where S is the observed MDR SST for each year in the period 1950-2005. The values of α and β are derived from model 4 in table 7 in Binter et al. (2007), in which observed basin hurricane numbers were regressed onto observed SSTs using data for 1950-2005. They have values of 6.25 and 5, respectively.

The basin hurricane numbers we create by this method should contain roughly the same long-term SST driven variability as the observed basin hurricane numbers, but different numbers of hurricanes in the individual years. We say 'roughly' the same, because (a) the linear model we are using to relate SST to hurricane numbers is undoubtedly not exactly correct, although given the analysis in Binter et al. (2007) is certainly seems to be reasonable, and (b) the parameters of the linear model are only estimated.

2.2 Generating artificial landfalling hurricane numbers

Given the 10,000 sets of simulated basin hurricane numbers described above, we then create 10,000 sets of simulated *landfalling* hurricane numbers by applying the rule that each basin hurricane has a probability of 0.254 of making landfall (this value is taken from observed data for 1950-2005).

The landfalling hurricane numbers we create by this method should contain roughly the same long-term SST driven variability as the observed landfalling series, but different numbers of hurricane in the individual years. They should also contain roughly the right dependency structure between the number of hurricanes in the basin and the number at landfall (e.g. that years with more hurricanes in the basin will tend to have more hurricanes at landfall).

2.3 Making predictions

We now have 10,000 sets of 56 years of artificial data for basin and landfalling hurricanes. This data contains a realistic representation of the SST-driven variability of hurricane numbers, and of the dependency structure between the numbers of hurricanes in the basin and at landfall, but different actual numbers of hurricanes from the observations. We can consider this data as 10,000 realisations of what might have occurred over the last 56 years, had the SSTs been the same, but the evolution of the atmosphere different. This data is a test-bed that can help us understand aspects of the predictability of landfalling hurricanes given SST.

The observed and simulated data is illustrated in figures 1 to 5. Figure 1 shows the observed basin data (solid black line) and the observed landfall data (solid grey line). The dashed black line shows the variability in the observed basin data that is explained using SSTs. The dotted grey line shows the variability in the observed landfall data that is explained using SSTs using the direct method, and the dotted grey line shows the variability in the landfall data that is explained using SSTs using the indirect method.

Figures 2 to 5 show 4 realisations of the simulated data. In each figure the dotted and dashed lines are the same as in figure 1, and show the SST driven signal. The solid black line then shows the simulated basin hurricane numbers and the solid grey line shows the simulated landfalling hurricane numbers.

We test predictions of landfalling hurricane numbers using the direct method as follows:

- we loop through the 10,000 sets of simulated landfalling hurricanes
- for each set, we miss out one of the 56 years
- using the other 55 years in that set, we build a linear regression model between SST and landfalling hurricane numbers
- we then use that fitted model to predict the number of landfalling hurricanes in the missed year, given the SST for that year
- we calculate the error for that prediction
- we then repeat for all 10,000 sets (missing out a different year each time)
- this gives us 10,000 prediction errors, from which we calculate the RMSE

We test the indirect method in almost exactly the same way, except that this time we also fit a model for predicting landfalling numbers from basin numbers.

2.4 Comparing the predictions

We compare the direct and indirect predictions in two ways:

- First, we compare the two RMSE values
- Second, we count what proportion of the time the errors from the direct method are smaller than the errors from the indirect method

We also repeat the entire calculation a number of times as a rough way to evaluate the convergence of our results.

3 Simulation-based analysis: results

We now present the results from our simulation study. The RMSE for the direct method is 1.61 hurricanes, while the RMSE for the indirect method is 1.58 hurricanes. This difference is small, but the sign of it does appear to be real: when we repeat the whole experiment a number of times, we always find that the indirect method beats the direct method.

The indirect method beats the direct method 51.8% of the time.

Given the design of the experiment, these results tell us how the two methods perform, on average over the whole range of SST values. Next year's SST, however, is likely to be warm relative to historical SSTs. We therefore also consider the more specific question of how the methods are likely to perform for given warm SSTs. Based on Laepple et al. (2007), we fit a linear trend to the historical SSTs, and extrapolate this trend out to 2011. This then gives SST values that are warmer than anything experienced in history (27.987^oC to be precise). We then repeat the whole analysis for predictions for this warm SST only. The results are more or less as before: the indirect method still wins, only this time by a slightly larger margin. The ratio of RMSE scores (direct divided by indirect) increases from 1.02 to 1.04.

4 The Linear normal case

We now study a slightly less realistic model, in which we take SSTs and hurricane numbers in the basin and at landfall to be normally distributed. These changes allow us to derive a very general result for the relative performance of the direct and indirect methods.

4.1 The setup

Here's how we set the problem up in this case.

Consider two simple regression models for centred random variables Y and Z,

$$\begin{aligned} Y &= X\beta + \varepsilon, \quad \varepsilon \sim (0, \sigma_\varepsilon^2 I_n), \\ Z &= Y\gamma + \eta, \quad \eta \sim (0, \sigma_\eta^2 I_n), \end{aligned}$$

where ε and η are independent. Here X, Y, Z, ε and η are $n \times 1$ column vectors, β and γ are scalars, and I_n is the $n \times n$ identity matrix. We will assume X is fixed.

In relation to the hurricane problem, X is the time-series of n years of SST values, Y is the time-series of n years of basin hurricane numbers and Z is the time-series of n years of landfalling hurricane numbers. Note that in our notation X is the *whole time-series* of SST, written as a vector, and similarly for Y and Z. Using vector notation avoids the messy use of subscripts. Two immediate comments about this setup: (a) we are assuming that basin and landfalling hurricane numbers are normally distributed. This doesn't really make sense, since they are counts that can only take integer values: using a poisson distribution would make more sense. We are starting off by addressing this question for normally distributed data because it's more tractable that way; (b) we are assuming a linear relationship (with offset and slope) between basin hurricanes and landfalling hurricanes. This is also a little odd, since there is no reason to have an offset in this relationship: if there aren't any basin hurricanes, there can't be any landfalling hurricanes. The most obvious model would be that each hurricane has a constant proportion of making landfall. Again, we are starting off by addressing this question in a linear context because it's more tractable that way.

We want to know about the accuracy of forecasts that we might make with the direct and indirect

methods. This translates mathematically into saying that we want to estimate

$$\begin{aligned} E(z_{n+1}) &= E(y_{n+1})\gamma && (2)\\ &= x_{n+1}\beta\gamma && (3)\\ &= x_{n+1}\delta && (4) \end{aligned}$$

where $\delta = \beta\gamma$.

The problem then boils down to measuring the quality of the estimator of δ since, if $\hat{z}_{n+1} = x_{n+1}\hat{\delta}$ is an estimator of $E(z_{n+1})$ then

$$\begin{aligned} \mathrm{MSE}(\hat{z}_{n+1}) &= \mathrm{MSE}(x_{n+1}\hat{\delta}) && (5)\\ &= E[(x_{n+1}\hat{\delta} - x_{n+1}\delta)(x_{n+1}\hat{\delta} - x_{n+1}\delta)'] && (6)\\ &= x_{n+1}\mathrm{MSE}(\hat{\delta})x'_{n+1}. && (7) \end{aligned}$$

So we now consider the direct and indirect methods for estimating δ.

4.2 Direct estimator of δ

We start by considering the direct, or one-step, method. This means we consider the relationship between X and Z, ignoring Y. The usual OLS estimator for δ is

$$\begin{aligned} \delta^{\dagger} &= (X'X)^{-1}X'Z && (8)\\ &= (X'X)^{-1}X'(X\beta\gamma + \varepsilon\gamma + \eta) && (9)\\ &= \delta + (X'X)^{-1}X'(\varepsilon\gamma + \eta). && (10) \end{aligned}$$

What are the statistical properties of this estimator?

In terms of mean:

$$E(\delta^{\dagger}) = \delta \quad (11)$$

i.e. the estimator is unbiased.

In terms of variance

$$\mathrm{Var}(\delta^{\dagger}) = (X'X)^{-1}X'\mathrm{Var}(\varepsilon\gamma + \eta)X(X'X)^{-1}. \quad (12)$$

We know that $\mathrm{Var}(\varepsilon\gamma + \eta) = \sigma_{\varepsilon}^2 I_n \gamma^2 + \sigma_{\eta}^2 I_n$, so

$$\mathrm{Var}(\delta^{\dagger}) = (X'X)^{-1}(\sigma_{\varepsilon}^2\gamma^2 + \sigma_{\eta}^2). \quad (13)$$

By equation 7 this then gives us an expression for the performance of the direct method.

4.3 Indirect estimator of δ

We now consider the indirect, or two-step, method. This means considering the relationships between X and Y, and Y and Z.

First, we consider estimating each regression separately. The OLS estimators for the slopes in each case are:

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'Y && (14)\\ &= \beta + (X'X)^{-1}X'\varepsilon && (15)\\ \hat{\gamma} &= (Y'Y)^{-1}Y'Z && (16)\\ &= \gamma + (Y'Y)^{-1}Y'\eta && (17) \end{aligned}$$

We now put the two models together, to create a single regression model based on the separate estimates for the two steps. We call the estimate of the slope of this combined model $\hat{\delta}$. Combining the expressions above, we have that:

$$\begin{aligned} \hat{\delta} &= \hat{\beta}\hat{\gamma} && (18) \\ &= \beta\gamma + (X'X)^{-1}X'\varepsilon\gamma + \beta(Y'Y)^{-1}Y'\eta + (X'X)^{-1}X'\varepsilon(Y'Y)^{-1}Y'\eta && (19) \end{aligned}$$

What are the statistical properties of this estimator $\hat{\delta}$?

It is clear (by independence of ε and η) that $\hat{\delta}$ is unbiased;

$$\begin{aligned} E(\hat{\delta}) &= \beta\gamma && (20) \\ &= \delta && (21) \end{aligned}$$

The variance is more awkward. Note that if ε were known then $\hat{\beta}$ and Y would be fixed constants. Thus,

$$\begin{aligned} E(\hat{\delta}|\varepsilon) &= E(\hat{\beta}\hat{\gamma}|\varepsilon) && (22) \\ &= \hat{\beta}E(\hat{\gamma}|\varepsilon) && (23) \\ &= \hat{\beta}\gamma, && (24) \\ \mathrm{Var}(\hat{\delta}|\varepsilon) &= \mathrm{Var}(\hat{\beta}\hat{\gamma}|\varepsilon) && (25) \\ &= \hat{\beta}\mathrm{Var}(\hat{\gamma}|\varepsilon)\hat{\beta}' && (26) \\ &= \hat{\beta}(Y'Y)^{-1}\hat{\beta}'\sigma_\eta^2. && (27) \end{aligned}$$

and so

$$\begin{aligned} \mathrm{Var}(\hat{\delta}) &= \mathrm{Var}(\hat{\beta}\hat{\gamma}) && (28) \\ &= E[\mathrm{Var}(\hat{\beta}\hat{\gamma}|\varepsilon)] + \mathrm{Var}[E(\hat{\beta}\hat{\gamma}|\varepsilon)] && (29) \\ &= E[\hat{\beta}(Y'Y)^{-1}\hat{\beta}']\sigma_\eta^2 + \gamma\mathrm{Var}(\hat{\beta})\gamma'. && (30) \end{aligned}$$

where we have used a standard relation for disaggregating the variance:

$$\mathrm{var}(a) = E[\mathrm{var}(a|b)] + \mathrm{var}[E(a|b)] \quad (31)$$

Using the facts that

$$\begin{aligned} E(Y'Y) &= \beta'X'X\beta + n\sigma_\varepsilon^2 && (32) \\ E(\hat{\beta}\hat{\beta}') &= \beta\beta' + (X'X)^{-1}\sigma_\varepsilon^2 && (33) \end{aligned}$$

and approximating to second order:

$$\mathrm{Var}(\hat{\delta}) = \left[\frac{\beta^2 + q^2}{\beta^2 + nq^2}\right](X'X)^{-1}\sigma_\eta^2 + q^2\gamma^2. \quad (34)$$

where $q^2 = (X'X)^{-1}\sigma_\varepsilon^2$.

4.4 Comparing the two estimators

We are now in a position to compare the estimators for the direct and indirect methods. Subtracting equation 34 from equation 13 gives:

$$\begin{aligned}
\mathrm{Var}(\delta^{\dagger}) - \mathrm{Var}(\hat{\delta}) &= (X'X)^{-1}(\sigma_{\varepsilon}^2\gamma^2 + \sigma_{\eta}^2) - \left[\frac{\beta^2+q^2}{\beta^2+nq^2}\right](X'X)^{-1}\sigma_{\eta}^2 - (X'X)^{-1}\sigma_{\varepsilon}^2\gamma^2 && (35)\\
&= (X'X)^{-1}\sigma_{\eta}^2 - \left[\frac{\beta^2+q^2}{\beta^2+nq^2}\right](X'X)^{-1}\sigma_{\eta}^2 && (36)\\
&= \left(1 - \left[\frac{\beta^2+q^2}{\beta^2+nq^2}\right]\right)(X'X)^{-1}\sigma_{\eta}^2 && (37)\\
&= \left[\frac{(n-1)q^2}{\beta^2+nq^2}\right](X'X)^{-1}\sigma_{\eta}^2 && (38)
\end{aligned}$$

The right hand side of this equation is clearly positive for $n > 1$.

This indicates:

- that using the indirect method is an improvement on the direct method, at least up to our second order approximations
- that if $\frac{\beta^2}{q^2}$ is small or σ_{η}^2 large then using the indirect method provides a marked improvement over the direct approach

5 Conclusions

We have compared the likely performance of direct and indirect methods for predicting landfalling hurricane numbers from SST. The direct method is based on building a linear regression model directly from SST to landfalling hurricane numbers. The indirect method is based on building a regression model from SST to basin numbers, and then predicting landfalling numbers from basin numbers using a constant proportion.

First, we compare these two methods in the context of a reasonably realistic model, using simulations. We find that the indirect method is better than the direct method, but that the difference is small.

Secondly, we compare the two methods in the context of a less realistic model in which all variables are normally distributed. For this model we are able to derive the interesting general result that the indirect method should *always* be better.

Which method should we then use in practice? If we had to chose one method, our results seem to imply that we should choose the indirect method, since it is more accurate. The simulation results suggest, however, that the performance of the two methods is likely to be very close for the values of the parameters appropriate for hurricanes in the real world. Given the possibility to use two methods we would use both, as alterative points of view.

Ideally we would also be able to solve the more realistic model analytically, as we have done for the linear-normal case. We are working on that.

References

R Binter, S Jewson, and S Khare. Statistical modelling of the relationship between Main Development Region Sea Surface Temperature and Atlantic Basin hurricane numbers. *arXiv:physics/0701170*, 2007. RMS Internal Report E04a.

T Laepple, S Jewson, J Meagher, A O'Shay, and J Penzer. Five-year ahead prediction of Sea Surface Temperature in the Tropical Atlantic: a comparison of simple statistical methods. *arXiv:physics/0701162*, 2007.

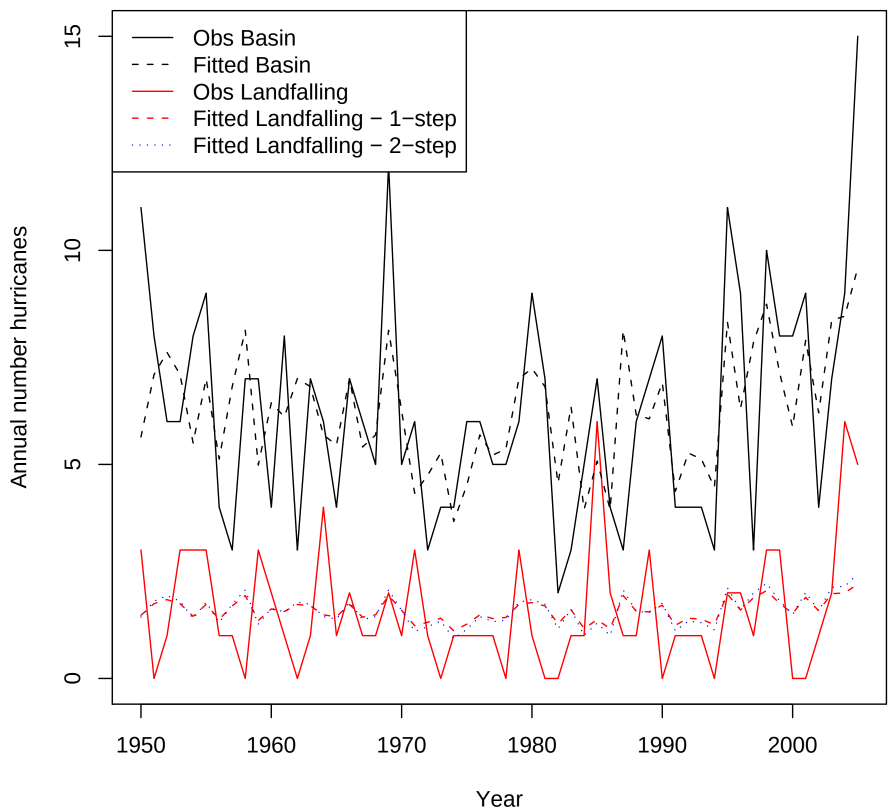


Figure 1: Atlantic basin and landfalling hurricane numbers for the period 1950 to 2005 (solid lines), with the component of the variability that can be explained by SSTs (broken lines).

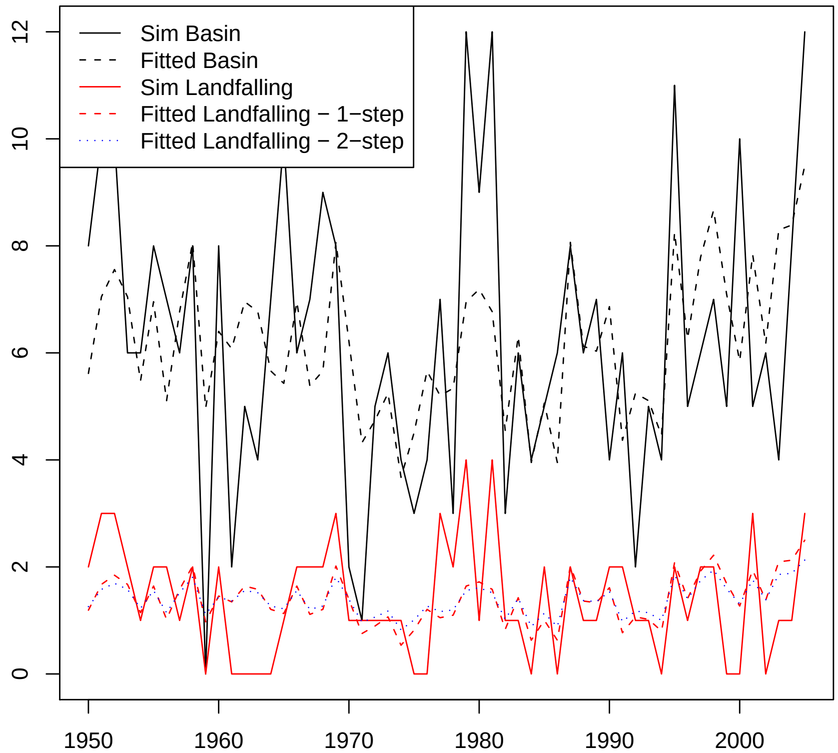


Figure 2: One realisation of simulated basin and landfalling hurricane numbers (solid lines), with the SST driven components (broken lines).

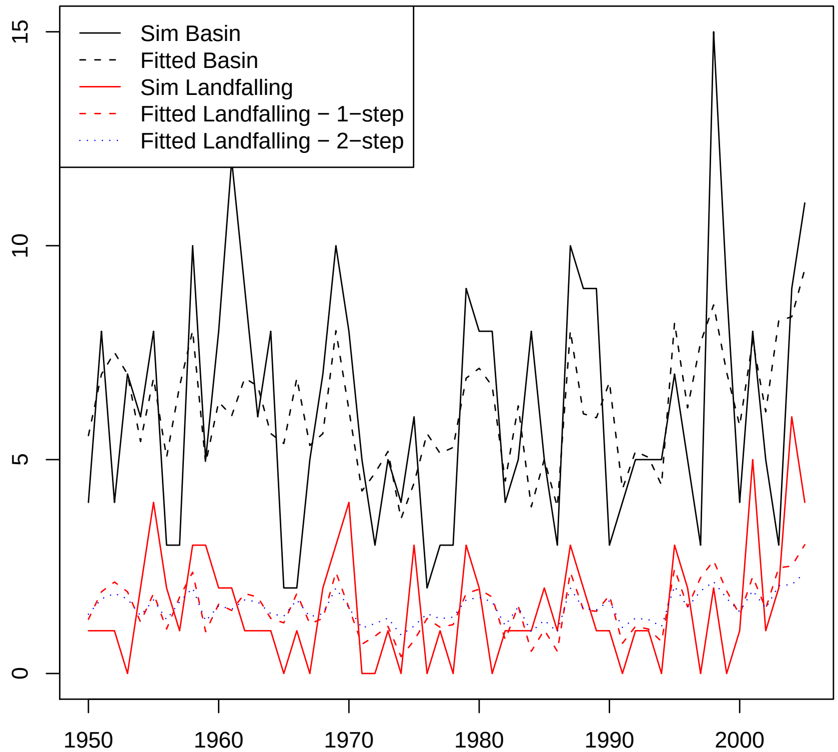


Figure 3: As in figure 2, but for a different realisation.

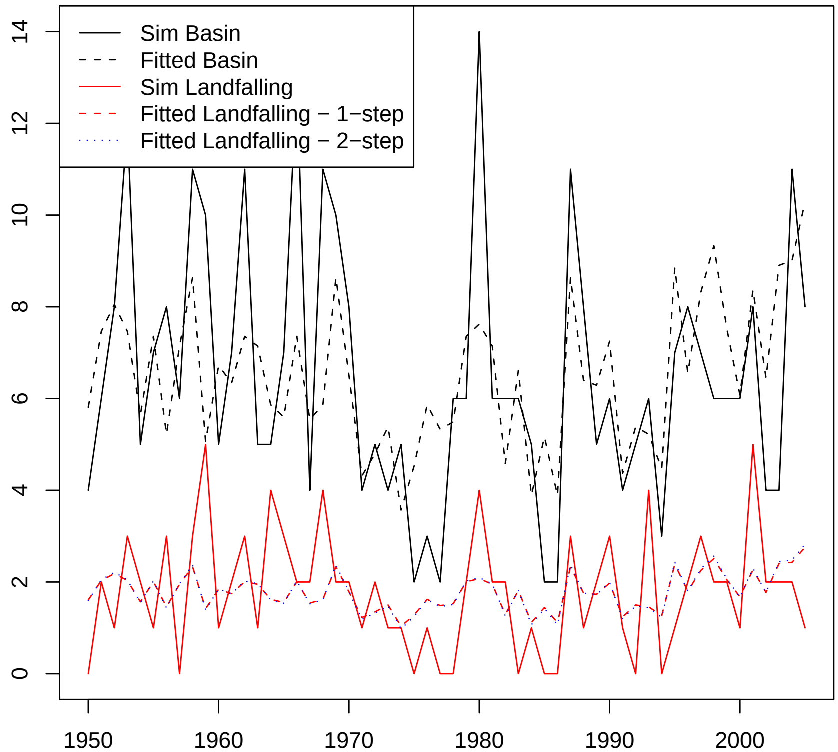


Figure 4: As in figure 2, but for a different realisation.

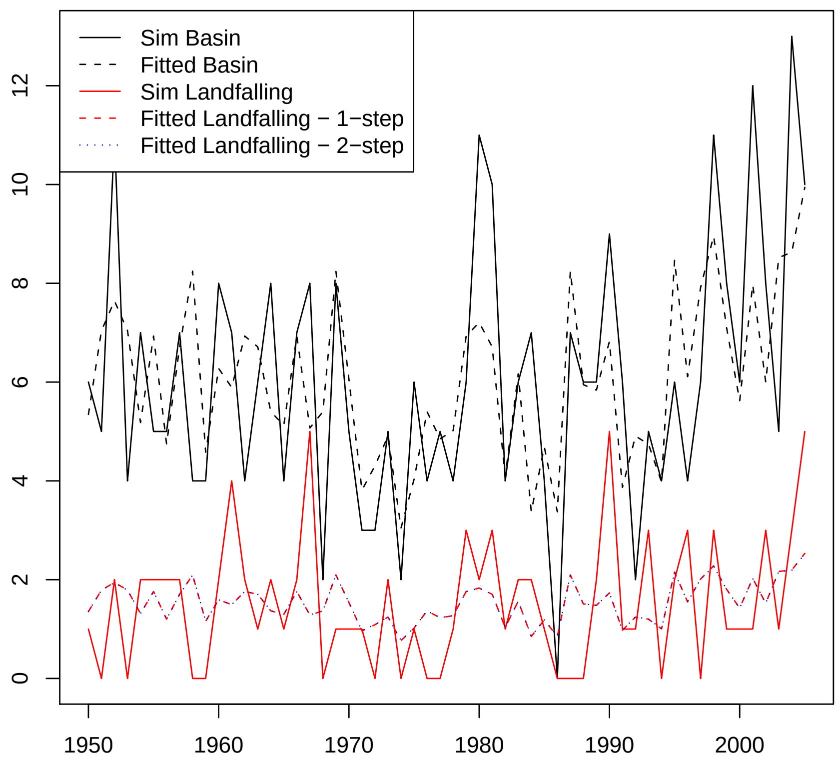
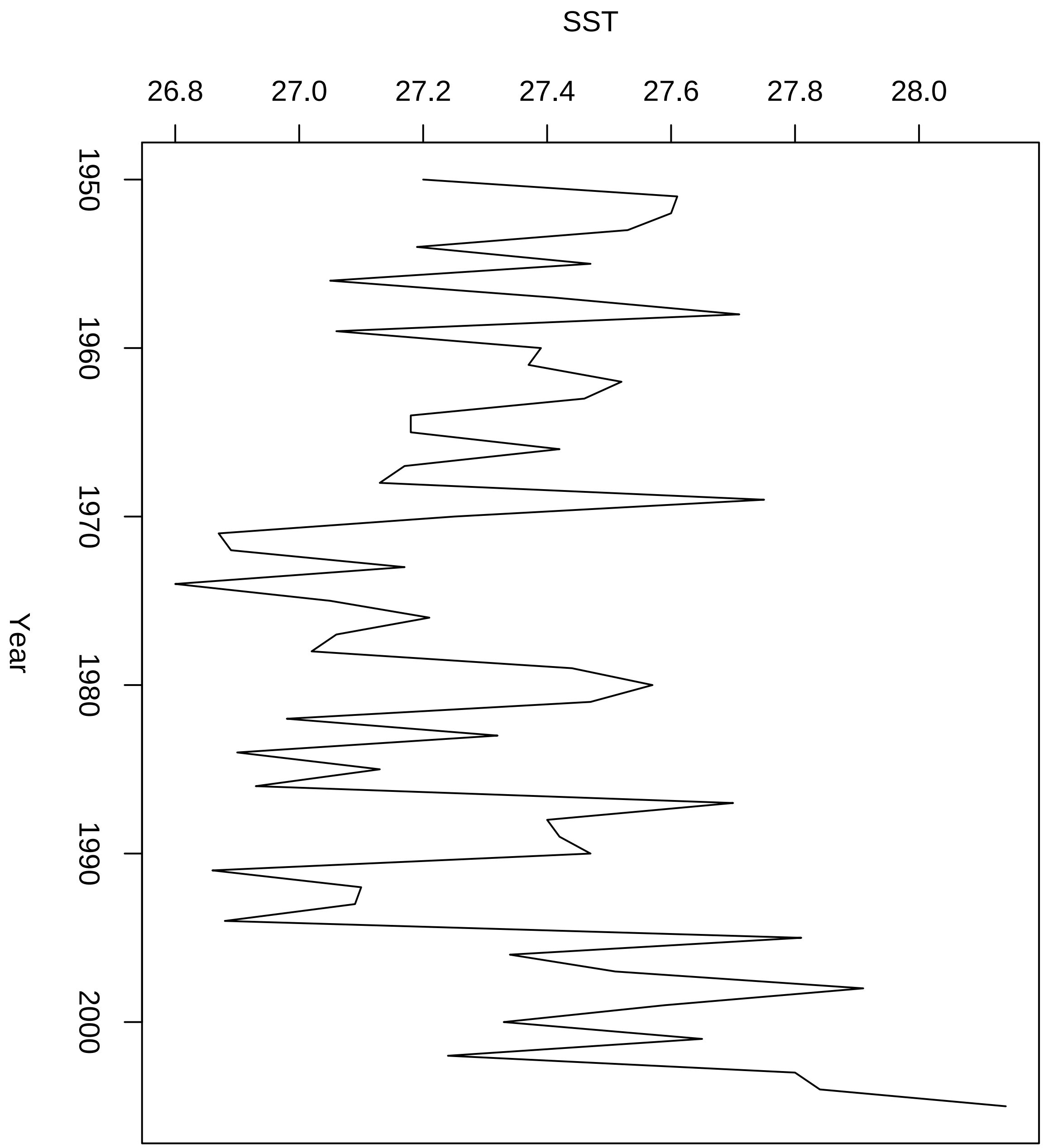


Figure 5: As in figure 2, but for a different realisation.


SST
26.8
27.0
27.2
27.4
27.6
27.8
28.0
Year
1950
1960
1970
1980
1990
2000